DOCUMENT RESUME

ED 152 851 TM 007 074

AUTHOR van der Flier, Henk; Drenth, Pieter J. D.
TITLE Fair Selection and Comparability of Testscores.
PUB DATE Jun 77
NOTE 21p.; Paper presented at the International Symposium on Educational Testing (3rd, Leiden, the Netherlands, June 27-30, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Admission Criteria; *Comparative Testing; Criterion Referenced Tests; *Cross Cultural Studies; Cultural Factors; Decision Making; Educational Testing; Elementary Secondary Education; *Equated Scores; Foreign Countries; Item Analysis; Mathematical Models; Minority Groups; Predictive Ability (Testing); Research Problems; Sex Differences; Socioeconomic Influences; Statistical Analysis; *Test Bias; *Test Interpretation; Test Validity
IDENTIFIERS *Construct Validity; *Decision Theoretic Models; Surinam

ABSTRACT

The criterion-oriented problem of test bias and fairness in selection is compared to the construct-oriented problem of comparability of test scores in cross-cultural research. These problems are shown to have important similarities, and their studies may supplement each other. In a formulation of psychometric criteria for the comparability of test scores, three levels are described: functional equivalence, score equivalence, and item equivalence. Patterns of item responses are considered to be useful for establishing the similarity of an individual to a group. Recent attitudes toward test fairness include a shift from criterion orientation to decision orientation, and growing doubt about the possibility of finding a generally acceptable objective definition of fairness. An experiment with two ethnic groups in Surinam using 12 tests and four background variables showed fewer differences related to ethnic group than to sex or to socioeconomic differences. The selection procedures for secondary school students in Tanzania are briefly compared to current models, and bibliographical references are appended. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152851

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Pieter J. D. Drenth

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

FAIR SELECTION AND COMPARABILITY OF TESTSCORES

Henk van der Flier and Pieter J.D. Drenth
Free University of Amsterdam

Third international symposium on educational testing,
Leiden, the Netherlands
June 27-30, 1977.

TM007 074

Fair selection and comparability of test scores.

H. v.d. Flier and P.J.D. Drenth.

During the past 15 years, the problem of test bias and fairness in selection has become an important subject in test psychology, especially in the U.S. Why particularly in the U.S.? We think there are a number of reasons:
In the first place, in the U.S.A. psychological tests play a more important role in the selection of individuals for all kinds of functions and educational programmes than in most other countries. This applies more so where the tests are mostly used in a formal, actuarial way, as in the U.S.A., and no, perhaps unconscious, correction for unfairness occurs, as does in a more subjective "clinical" approach.
In the second place, in the U.S., one is dealing with relatively large racial and cultural minority groups who, from a social economicas well as educational point of view, relatively often belong to the lowest levels of the population.
In the third place finally, there is the ideology of "equal opportunities for individual advancement" which has occupied a dominant position in the U.S. in the past decades.

Until now, the problem has not received such extensive attention in Europe, or in our own country, (Holland), either, but that appears to be changing, especially regarding the increased interest in the influence of environmental factors on school choice and school achievements, and also as a result of the development of quite considerable cultural-racial minority groups, Consider, for instance, the various groups of foreign workers from Southern Europe and North Africa as well as the large group of people from Surinam* who emigrated to Holland shortly before Surinam gained independence in 1975. However, as yet, we have not come much further than identifying the problem and little systematic research has been carried out.
A problem that has received much more attention in Holland, and one that especially plays an important role in cross-cultural research, is the comparability of test scores. The issue here is wether, and to what extent, the same qualitative and quantitative meaning can be attached to test scores in different cultural groups. Thus, when two groups widely differ in their scores on a spatial reasoning test, the question is, to what extent these differences do reflect actual differences in spatial reasoning,

* An ex colony of the Netherlands situated at the Northern coast of South America.

and to what extent they are, for example, only the result of differences in adjustment to the artificial test situation, or in other culturally or educationally determined skills necessary for the solution of the test problems.

Strangely enough, although the problems of comparability and fairness display many similarities, and the analysis techniques applied are similar, the two areas have remained rather separate. To an extent, there seems to be quite a practical reason for this: with fairness, we are always concerned with groups that function in one and the same social setting (applicants for the same job, or the same educational programme) which is not the case, or to a far lesser degree with cross-cultural research. However, there is also an important difference in theoretical orientation. The requirement of comparability belongs in a construct oriented approach. One determines whether the test refers to a particular construct in a similar way in the distinguished groups, and the criteria used are generally derived from test theory. Fairness, on the other hand, belongs within a criterion- or decision oriented approach. A test is said to be biased if the differences in testscores are too large or too small compared with the differences on some criterion measure for which the test is used as a predictor, or if the proportions of the groups that are selected or rejected are too large or too small compared with the proportions of both groups that will be succesful on the job or in the educational program.

Without directly wanting to express a preference for one approach or the other, we feel that it would be useful to place them side by side and ascertain whether they can supplement each other. Firstly, (let us say) something about comparability: The meaning of the concept can best be clarified by using the distinction Vernon (1969) makes between intelligence A, B and C: Intelligence A is the genetically determined intelligence one receives from the parents at birth, or in fact 9 months prior to birth. Intelligence B is the phenotypical aspect of intelligence. It refers to the cognitive abilities; being measured and which are dependent on the interaction between ability (Intelligence A) and environment. Intelligence C finally, relates to the actual test results, expressed in operational terms. It is, of course, directly dependant on the chosen measuring instrument. The symposium on Heredity of Intelligence and Education dealt with the relationship between A and B - comparability research is concerned with the relationship between B and C.

The relationship between A and C has, in the past, also played a role in cross-cultural research - in particular, refering to certain tests as culture-free, raised expectations that the influence of environmental factors might have been eliminated, and that is might be possible to conclude differences in ability between different cultural groups on the basis of the test scores. However, attempts to construct culture-free tests have never succeeded. This is not surprising when we consider that, in this line of thought, even the ordinary ruler would not be a culture-free instrument as long as the length measured was partly dependent on environmental factors.

Now back to comparability. So the question is, whether the test scores (level C) similarly represent the attribute (B) in different cultural groups. To be able to answer this question, the point of departure must be that the attributes or constructs in question have general validity. Therefore, it must be assumed that traits like spatial reasoning and verbal ability are universal (the so called etic approach) and appear in all groups, although the manifestation of those traits can, of course, be culture specific. Naturally, this assumption is disputable and there are a number of researchers who feel that direct comparison is impossible, and that cognitive functioning can only be understood as adjustment to specific cultural and ecological conditions (the emic approach). For a survey of this discussion, we refer you to publications by Frijda and Jahoda (1966), Eckensberger(1971), Berry (1972), Poortinga (1975), and Jahoda (1977). We will confine ourselves here to the approach in which an attempt is made to formulate psychometric criteria for the comparability of test scores, with the universality of constructs as point of departure.

One of the most elaborate approaches in this field is Poortinga's (1971, 1975), which disthinguishes between different levels of comparability. These are functional equivalence, score equivalence and item equivalence. Functional equivalence relates to the requirement that the test qualitatively measures the same attribute in the different groups. When dealing with more than one test to which this requirement applies, the testable condition follows, that the correlations between the different testsshould be similar in the groups concerned. Thus it is tested whether <u>both</u> instruments are functionally equivalent. Only when the comparability of one of the tests can be assumed, or when dealing with comparable reference variables, can statements be made about single tests. Score equivalence implies that the test should measure the attribute on a similar scale in the distinguished groups;

when there is more than one test, the requirement of similar regression coefficients follows. Thus, here too, the score equivalence of two tests is studied at the same time - unless it can be assumed that one of the two variables is score equivalent. Item equivalence finally, implies the same requirements as with score equivalence but then, relatingto the individual items. The condition that can be derived from this, is that the items should be similarly related in the different groups, and also that, under certain conditions, the sequence of the p-values should be similar.

Some critical notes may be appropriate here:

I) To be able to propose, as a condition for functional equivalence, that the inter-correlations of the tests, or the correlations with the reference variables should be the same in different populations, it must be assumed that the theoretical attributes the tests are supposed to measure are similarly related in these populations; likewise, when testing score equivalence, it is assumed that the regressions of the theoretical attributes, or at least the regression coefficients, are similar. In general, little can be said about the plausibility of these assumptions, which clearly imply more than just universality of constructs - namely, the universality of nomological networks in which the construct has it's place. However, it is clear that there is a danger that the conditions for the comparability of test scores, provided by this method, are too stringent. For instance, to label the scores on a ruler and scales as incomparable when a different relationship between length and weigth is found in two racial groups seems hard to defend.

2) From the definition of score equivalence, one would expect the testable condition that the regressions of the tests on each other should be similar in the populations concerned. However, these conditions can only be met when the mean testscores are the same in both populations. For this reason, testing the similarity of the regression coefficients, is considered to be sufficient and the intercept is not taken into account. This solution however, is rather unsatisfactory because the intercept really is important when we are dealing with quantitative comparability of test scores. After all, the question is - to put it simply - whether group (culture) A has more of a particular attribute than group (culture) B.

3) In our opinion, the condition of item equivalence offers the best possibilities from a theoretical point of view for comparability research, because we are dealing here with indicators of one and the same construct; certainly when the test is reasonably homogenous.

Generally speaking, the acceptability of a particular construct depends on the internal consistency of the set of variables which are considered to represent it, and also on the relationships between these variables and variables which, according to the description of the construct, do not belong to the set. The relative invariance of all these relationships in a cultural group is a prerequisite for cross cultural comparison. When the network of relationships also shows cross cultural constancy, it can be assumed that the set of variables refers to the same construct in both groups. The more systematically the construct is operationalised, and the more the representing variables differ in nature, the more this applies*.

The simplest of the conditions for item equivalence to apply, is the similar sequence of p-values. Irvine (1968) used this condition in his study of the comparability of the Progressive Matrices Test in different parts of Africa, and Poortinga has also used it frequently, among other occasions, in a study of the comparability of a number of curiosity tests for blacks and whites in S. Africa (Poortinga and Foden, 1975). However, two serious problems arise regarding this condition: in the first place, the condition of similarity in the sequence of p-values is only justified when the score distributions in the groups under comparison are the same, or when the trace lines of the items do not cross each other. In the second place, the hypothesis that the rank correlation between the series of p-values, corrected for attenuation, is one is, by definition, difficult to test, while the obvious alternative-testing of the interaction between items and group membership also encounters difficulties, considering that we are dealing with a dichotomous dependent variable.

The condition of similar intercorrelations of the items in the two groups has its problems as well. Correlations between (dichotomous) items partly depend on the difficulty-level of the items, and so, strictly speaking, this condition too is only justified when the score distributions are similar in the two groups.

The principal objection to the conditions of item equivalence discussed is that they have an ad hoc character. It seems better to explicitly start from a particular scaling model when specifying the parameters from which the similarity must be tested.

* Of course, this argument also holds for a set of tests which are considered to represent one and the same factor.

One problem with this of course, is that models which offer a lot of guarantees for comparability also make strong demands on the data. For instance, Mokken (1970) studied the comparability of a political efficacy scale in the U.S. and Holland applying a probabilistic variant of the Guttman model, and comparing the scalability indices of the items and the total sets in the two groups. The so called Mokken scale however, presupposes holomorphic sets of items, which implies that the trace lines of the items should not cross each other. It is doubtful whether this condition is realistic for most ability tests. The same applies to the Rasch model. Mellenbergh (1972) made a study of the comparability of a mathematics test with Dutch and Indonesian school children. The aim of the study was to compare the item parameters in the two groups (according to the Rasch model). However, the data did not meet the requirements of the model in either of the two groups. Possibly, the more general latent trait models will provide a more important contribution to comparability research. Lord (1977) describes a procedure for testing the similarity of item characteristic curves for different samples where less stringent restrictions are imposed on the data.

All these approaches mentioned confine themselves to groups that are defined beforehand. However, it is also possible to apply the requirement of item equivalence on an individual level and, on the basis of the results, to form groups in which the scores have a high degree of comparability. This procedure has already been more fully discussed elsewhere (v.d. Flier 1977) so we will only treat it briefly here: essentially, it involves determining the degree in which someone's score pattern on the items, deviates from an expected score pattern. The degree of deviation is expressed in a so called deviance score. The expected score pattern can be defined a-priori, or can be determined from the mean scores in a particular group. With dichotomous items, these are the p-values. Suppose now that the items in a test are placed in order of difficulty (from easy to difficult). It is then expected that someone who correctly answers x of the n items, will answer the first x items correctly and the last $n-x$ items incorrectly. Should he answer the easy items incorrectly, then his score pattern is regarded as extremely deviant. The deviance score is intended to indicate someone's position between these two extremes. So the deviance score reflects the degree in which the items have the same order of difficulty for the testee as for the reference group.

In principle, the calculation of deviance scores is not very complicated: firstly, the sum of the rank numbers of the x correct items is calculated, from which the sum of the rank numbers of the x easiest items is subtracted. The difference is then divided by the difference between the sum of the rank numbers of the x most difficult items and the sum of the rank numbers of the x easiest items. Thus, for example, when we have a set of five items, which are placed in order of difficulty, and we find the following score pattern (IOIIO), the deviance score is then equal (8-6)/(12-6)=.33. Some technical complications must be kept in mind. In the first place, the order of p-values of the items may change with ability level, so the p-values and the deviance scores will have to be separately determined for different score groups. In the second place, it is also necessary to take the distances between the p-values in the set into account, to obtain a good chance distribution of deviance scores. This can be accomplished by, as it were, bedding the items in a fixed superset of items and carrying out the calculations on the basis of the rank numbers they would get in the superset.
We feel that the application of deviance scores, may offer better possibilities for differentiated insight, and is more closely corresponding to the more recent developments in test psychology. By not starting from a-priori defined groups but by letting an optimal differentation develop, as it were, from an analysis of the material itself, a more differentiated description seems possible. This approach is also allied to recent attempts to give the qualitive aspects of the test achievements a more central place, or in other words, to the switch from the test score to the process behind the score. (Cronbach and Drenth, 1972, p. 472).

The recent developments in thought about fairness in selection are characterised by two aspects: in the first place, a shift from a criterion orientation towards a decision orientation, and in the second place, increasing doubt as to the possibility of arriving at a generally accepted objective definition of fairness. The classic regression model as described by Cleary (1968) and others, in which fairness is more or less equated with optimal prediction, and the corresponding Equal Risk Model of Einhorn and Bass (1971) which has a dichotomous utility scale as point of departure, could admittedly be called fair with regard to the separate individuals but the result still appeared to be that the relative proportion of sollicitants from the minority group that would be accepted on the basis of the test scores would, in many cases, be smaller than the proportion that could attain a specified level of criterion achievement. Moreover, unfairness in terms of these definitions arose less frequently than was expected, Cronbach (1972) even suggested that the

findings more frequently indicated bias against the majority group. So the basic social problem behind the concern with fairness - namely, that minority groups got relatively fewer chances of sharing socially desirable facilities or opportunities(education, work) did not appear to be solvable with the aid of these definitions.
Thorndike (1971) and Coles' (1973) models can be regarded as attempts to achieve more fairness at group level. We will not examine these models in detail as Petersen has already dealt them comprehensively in her paper, and we suffice with the observation that fairness here goes at the cost of optimal prediction.
Darlington (1971) was probably one of the first authors to drop the idea of an objective definition of fairness. He proposes that when the aims of maximal validity and maximal equality of chances conflict, a subjective decision should be made regarding the relative importance of the two aims. His proposition ammounts to letting the decision maker indicate how much of the desired criterion output could be offered up to increase the minority group's chances of being accepted; or in other words, the difference in criterion achievement at which solicitants from the majority and minority group can be regarded as equally acceptable.
So a certain value is added to the criterion score of members of the minority group, and this corrected criterion is subsequently predicted as well as possible.
This line of thought, in which the subjective value judgement is assigned a distinct place, is further generalised in the decision theoretical approaches of Gross and Su (1975) and Petersen and Novick (1976). Subjective utility values, which can differ for the majority group and the minority group, are awarded to the various possible outcomes of the selection process. Cutt-off scores for the two groups are then derived on the basis of the chances of the different outcomes in the two groups, and the utilities awarded - in such a way, that the expected utility of the selection process is maximized.

In our opinion, this development towards a decision theoretical approach should be positively valued - especially because, in this approach, the criterion occupies a less absolute place, and also because it allows more room for contribution from test theory and comparability research than the classic approach. We will now examine these points in detail:

1. The criterion oriented approach as embodied by the classic regression model, relies heavily on the fairness of the criterion. The differences on the predictors that can be considered fair, are derived from the differences in criterion achievement between the distinguished groups.

So the model is almost by definition, a conservative one in the sense that existing dissimilarities within the total system of predictors and criteria are continued and even amplified. This means that the model is not usable in those cases in which the intention is to break through the existing system. The expected utility model on the other hand, does allow social and political value judgements falling outside the system to be included in a rational way in the selection decisions. For that matter, in this approach too, the starting point is that the criterion measure, on the basis of which the outcomes of the selection process are specified, is "fair".
It would therefore seem sensible to study criterion measures for bias as well before using them. For example the conditions developed in comparability research (functional equivalance, item equivalence) could be applied. It is quite possible that the population we lump together at present with respect to criterion achievement is not so homogenous at all, and can not be treated as such. An analysis such as this might then lead to differentiations of the criterion and of the population for which predictions have to be made, which would link up with an argument Dunnette already made some time ago, for a "modified model" for test validation and selection research. (Dunnette, 1963).
Finally, in those cases in which there are no unbiased external criteria available it will also be preferable to apply one of the various comparability checks in which the predictor - criterion relations are not taken into consideration.

2. The expected utility model in no way requires the conditional probabilities to be separately determined for the group for which different utility values are specified. The different utilities are awarded on the basis of social considerations, while the determination of the chances of the various outcomes is an empirical matter where all that counts in fact, is that the moderator showing the greatest differences in test-criterion relationships is the best.
So, although with regard to the utilities a distinction between more or less apriori distinguished cultural/racial groups is perhaps the most obvious, this does not also have to apply to the specification of the conditional chances.

The emprical evidence for differential validities with regard to racial minority groups is not very impressive, as is apparent from the reviews of Schmidt et al. (1973), O'Connor et al. (1975), Katzell and Dyer (1977), Boehm (1977) and Bartlett et al. (1977).
Differences in regression coefficients and intercepts are more often found (Kirkpatric et al., 1968, Gael et al., 1975), but these differences are not very consistent, and their size hardly justifies the greater ammount of work and the greater inaccuracy that accompanies working with separate groups.
This result is not really very surprising - membership of a cultural/racial group is a complex variable which in itself has little explanatory value regarding differences in test and criterium achievements. Or to quote Fincher (1975):
"Race is meaningful only in the sense that it is a proxy variable for socioeconomic status, education, occupational history, previous training, cultural deprivation, or other variables that may be regarded as determinants".
In our opionion - especially now that social and political value judgements relating to the emancipation of cultural/racial minority groups can be expressed by means of utilities - the research into moderator variables should be more strongly directed towards these direct determinants.
This point can be illustrated by the results of a test development project that was carried out in Surinam at the end of 1973 with almost 400 primary school children (v.d. Flier and Marshall 1974). Ten different tests were administered in total - all non-verbal. In addition information was gathered about a number of background variables including the cultural/racial group, sex, quality of the primary school, linguistic background, and the father's occupation. Finally, the exam results for the two most important school subjects Maths and Language were collected at the end of the school year. 300 of the pupils took part in the exam. The question now is whether, and to what extent the relationships between the tests and criteria differ for the subgroups that can be distinguished on the basis of the background variables.
Regarding the variable cultural/racial group, a distinction can be made between Creoles, who originally came from W. Africa and who form the largest population group in Surinam and Aziatics who mainly originate from India and Indonesia.

The latter group is clearly more heterogenous than the former. We first give the means and standard deviations of the two sub-groups for the tests and the criteria.

Table I Means and standard deviations of Creoles and Aziatics for the tests and the criteria.

Tests/criterion-measures	Creoles (N=232/170)		Aziatics (N=181/124)		
	$\bar{X}$	SD	$\bar{X}$	SD	t
1. Mazes	9.00	4.63	8.17	5.58	n.s.
2. Exclusion	13.48	5.01	13.31	5.45	n.s.
3. Components	9.30	3.95	9.54	3.81	n.s.
4. Differences	17.41	5.30	18.48	4.41	-2.23*
5. Numbers	23.59	11.66	27.14	11.16	-3.35*
6. Triangles	34.16	13.66	33.87	13.88	n.s.
7. Hidden Patterns	55.16	24.19	53.98	26.62	n.s.
8. Coding	27.52	11.02	28.77	11.92	n.s.
9. Signs	204.93	55.27	208.49	56.02	n.s.
10. Groupstest 70/1	52.51	18.41	52.52	18.70	n.s.
Maths	4.71	2.21	5.58	2.05	-3.63**
Language	6.62	1.32	5.94	1.36	2.00*

* $p < .05$

** $p < .01$

The differences between the two groups are not particularly large. The group, Aziatics, has significantly higher scores for the Perceptual Speed test Differences, and for the the subtest Numbers, that measures facility in simple numerical computations. As far as the exams are concerned, the Aziatics are somewhat better in Maths while the Creoles are slightly better in Language. We will now look at the correlations between the tests and the criteria in the total group and in the two sub-groups.

Table II Correlations of the tests with Maths and Language for the total group, and for Creoles and Aziatics separatley.

Tests	Maths				Language			
	Total (N=294)	Creoles (N=170)	Aziatics (N=124)	Difference C - A	Total (N=294)	Creoles (N=170)	Aziatics (N=124)	Diff. C - A
1. Mazes	.21**	.22**	.23**	n.s.	.21**	.13*	.30**	n.s.
2. Exclusion	.20**	.25**	.16**	n.s.	.17**	.16*	.17*	n.s.
3. Components	.10*	.18**	-.04	n.s.	.13*	.15*	.11	n.s.
4. Differences	.22**	.25**	.13	n.s.	.26**	.28**	.26**	n.s.
5. Numbers	.68**	.69**	.62**	n.s.	.20**	.25**	.20*	n.s.
6. Triangles	.22**	.21**	.24**	n.s.	.13*	.10	.16*	n.s.
7. Hidden Patterns	.11*	.16*	.04	n.s.	.18**	.19**	.18*	n.s.
8. Coding	.35**	.38**	.30**	n.s.	.22**	.25**	.19*	n.s.
9. Sign	.08	.10	.02	n.s.	.16**	.17*	.15*	n.s.
10. Grouptest 70/1	.15**	.17*	.15*	n.s.	.11*	.08	.14	n.s.

* $p < .05$

** $p < .01$

Nearly all tests appear to be significantly related to the two criteria, but on the whole, the correlations are quite low. As to be expected, the exam mark for Maths seems best predicted by the subtest Numbers, and additionally, by the subtest Coding which also involves the manipulation of symbolic material. Differences, and Coding once again, correlate the highest with Language.

There are no great differences between the correlations in the two sub-groups- regarding several tests, one could speak of single group validity, but in none of the cases are the differences significant.

We will now proceed with table III, in which a summary is given of the differences in validities and regression equations with diverse possible divisions of the total group. Thus with each of these divisions, we are comparing 20 test-criterion relationships. Consecutively given are, the number of cases of single group validity (in which a significant relationship is found in one group but not in the other), the number of cases of differential validity (in which the correlations in the two groups differ significantly), the mean difference between the correlations in the two groups (after conversion into z-scores) and the number of cases in which the regression weights and the intercepts in the two groups significantly differ from each other.

Table III Comparison of validities and regression equations with different divisions - summary table.

	n single group validity	n differential validity	mean difference $\lvert Z_1 - Z_2 \rvert$	n different regression coefficients	n different intercepts
1. Ethnic group:					
Creole (N=170) - Asian (N=124)	5	0	.076	1	12
2. Sex:					
Boys (N=146) - Girls (N=148)	11	4	.128	5	4
3. School quality:					
High, medium (N=212) - Low (N=82)	15	3	.144	3	16
4. Linguistic background					
Dutch (N=217) Other (N=77)	17	4	.154	3	10
5. Fathers occupation					
White collar (N=103) Other (N=130)	12	8	.181	5	11

Of the five background variables, on the basis of which the group was divided, we have already discussed Ethnic group, and Sex needs no further amplification.

The classification of schools into high, medium and low quality schools is primarly based on the exam results in preceding years. The group of pupils who attend a school of medium quality is larger than the other two groups - which is why this variable is rather unevenly divided. That also holds for the variable linguistic background, which relates to the degree in which the language used in education (Dutch) is also used at home. Dutch is mainly spoken by Surinamers with a Creole background - which accounts for the Creole group's higher scores in the language exam.

The variable occupation of the father was divided into white collar jobs and skilled - and unskilled jobs. This provides a fairly even distribution. The correlation of "Fathers occupation" with school quality is .32, and with linguistic Background .35.

When looking at the results, it can be seen that the division into Ethnic groups yields the fewest differences in validities and regression coefficients of all divisions. The best moderator is "Fathers occupation".

As to be expected, the number of cases of single group validity is considerably higher than the number of cases of differential validity. This especially applies to the variables School Quality and Linguistic Background, where, because of the difference in size of the subgroups, cases of single group validity can relatively easily arise, and there is relatively less likelihood of cases of differential validity.

In all cases of differential validity, the differences in the relevant division are in the same direction. The validities are higher for girls than for boys, higher for pupils from high and medium quality schools than for pupils from low quality schools, higher for pupils who frequently speak Dutch at home, and especially higher for pupils with higher social-economic backgrounds.

The differences between the regression coefficients run more or less parallel to the differences between the correlation coefficients. The similarity of the coefficients was tested with a t-test which is described in Plomp (1972). The number of significant differences with the different divisions is given in the 4th column of the table. Only in those cases in which the regression coefficients did not differ significantly, the difference between the intercepts was tested. That was done with the help of an F-test, described in Kerlinger and Pedhazur (1973), which ascertaines whether the criterion is better predicted on the basis of the predictor and a variable that indicates group membership, than solely on the basis of the predictor. As can be seen in column five in the table, there appear to be quite a few differences in intercept with the various divisions. This also applies to the division into Creoles and Aziatics. However, in our opinion, not too much value can be attached to these differences because we are concerned here with tests that have predominantly low correlations with both criteria. If, in such a case, two subgroups differ in criterion performance, the difference on the predictor will have to be considerably larger that the difference on the criterion to obtain the same regression line. In fact, this was also Thorndike's criticism of the classic regression model (Thorndike, 1971).

Be that as it may, this effect will appear to a far lesser degree when the tests are selected on the basis of their predictive value with regard the criterion, as will mostly be the case in practice. So, it can be concluded with some reserve that divisions according to sex, school quality, linguistic background or social-economic level seem to show more result in the prediction of school success in Surinam than a division according to ethnic group.
One last brief remark on this point: besides dividing the group into subgroups on the grounds of individual background variables, or combinations of background variables, one could also consider a division based on differences in the way the test achievements came about. In this respect we are of course mainly thinking of the possibility of making a distinction, on the basis of deviance scores, between groups in which the test has more or less predictive power. At this moment however, there are no detailed results available about this.

3. Without doubt, the most important problem in applying the Expected Utility Model is the determination of the utilities. Gross and Su (1975) strongly emphasise this point, and come to the conclusion that, in view of the conflicting interests of the different groups involved it may not be possible to arrive at a selection procedure which is regarded as fair by all parties. Cronbach (1976) also devotes much attention to the conflicts of interests, especially between employers and applicants, but he does not offer very concrete solutions. A related problem is the danger of applying non-ethical utilities. Where the earlier "objective" definitions of fairness seemed to offer a certain guarantee against all too blatant a misuse of test results, the Expected Utility model explicitly allows room for subjective value judgements, and the principal guarantee against misuse seems to be the fact that utilities, at least with the current still quite simple pay-off matrices, lend themselves well for public discussion. Novick and Petersen (1976) have a fair ammount of confidence in the outcome of this discussion. Perhaps they are right as far as the U.S. is concerned, although it is still questionable whether more than equal utilities for minority groups would result in a real democratic decision. However, greater problems could be expected in situations in which the majority of the population declared itself to be a minority group, and the utilities would be used to reduce the possibilities of a minority group, rightly or wrongly considered as priveleged. This situation could easily arise, particularly in many developing countries.

Although, as mentioned, simple pay-off matrices have the advantage that they lend themselves well for public discussion they are also on the other hand not very realistic, Now of course it would be wrong to regard utilities as values to be derived from the desired result of the selection procedure. On the other hand discrepancies can also arise if a person's utilities can not be expressed in a simple system.
As an example let us take the selection for the secondary school in Tanzania. We will restrict ourselves to the official regulation which is, that from each of the 19 regions into which Tanzania is divided, an equal percentage of pupils from the 7th class of the primary school are admitted - inspite of the fact that there are large differences in exam results between these regions (that also widely differ in economic development). We can additionally mention that the selection ratio is extremely low (there are only places for 7% of primary school leavers in secondary schools).
Now the most important point is that the utilities are not fixed here. The outcome of the selection process in pre determined, at least as far as the regions are concerned, and the "ammount of succes" that is offered up for the sake of better chances for the under-developed regions varies with the empirical situation. That is with the size of the differences between the regions.
Although a selection procedure such as this is perhaps not logically consistent in terms of the current models, we can still not be too unhappy with it, and we can imagine that more complex models will fit more closely to this situation. In this respect, Darlington's (1976) suggestion to let the preference given to a minority applicant depend upon the number of minority applicants already selected, may be a good point of departure.

REFERENCES

Barlett, C.J., P. Bobko and S.M. Pine (1977). Single-group validity: Fallacy of the facts? J. Appl. Psychol. 62, 155-157.

Berry, J.W. (1972). Radical cultural relativism and the concept of intelligence, in Cronbach, L.J., and Drenth, P.J.D. Mental Tests and Cultural Adaptation. Den Haag, Mouton.

Boehm, V.R. (1977). Differential prediction: A methodological artifact? J. Appl. Psychol. 62, 146-154.

Clearly, T.A. (1968). Prediction of grades of negro and white students in integrated colleges. J. Educ. Meas. 8, 71-81.

Cole, N.S. (1973). Bias in selection. J. Educ. Meas. 10, 237-255.

Cronbach, L.J., and P.J.D. Drenth (1972). Mental tests and cultural adaptation. Den Haag, Mouton.

Cronbach, L.J. (1972). Judging how well a test measures. In: Cronbach, L.J. and P.J.D. Drenth (1972). Mental tests and cultural adaptation. Den Haag, Mouton.

Cronbach, L.J. (1976). Equity in selection - where psychometrics and political philosophy meet. J. Educ. Meas. 13, 31-41.

Darlington, R.B. (1971). Another look at "cultural fairness". J. Educ. Meas. 8, 71-82.

Darlington, R.B. (1976). A defense of rational personnel selection, and two new methods. J. Educ. Meas. 13, 43-52.

Dunnette, M.D. (1963). A modified model for test validation and selection research. J. Appl. Psychol. 47, 317-323.

Eckensberger, L. (1970). Methoden problemen der Kultur vergleichende Psychologie. Saarbrücken: D. Breitenbach Verlag.

Einhorn, H.J., and A.R. Bass (1971). Methodological considerations relevant to discrimination in employment testing. Psychol. Bull. 75, 261-9.

Fincher, C. (1975). Differential validity and test bias. Personnel Psychology, 28, 481-500.

Flier, H.van der, and E.K. Marshall (1974).Testdevelopment in Surinam. Amsterdam, Free University.

Flier, H. van der (1977). Environmental factors and deviant response patterns. In Y.H. Poortinga (ed.): Basic problems in cross-cultural psychology. Amsterdam, Swets & Zeitlinger, 30-35.

Prijda, N. and G. Jahoda (1966). On the scope and methods of cross-cultural research. Int. J. Psychol. 1, 109-127.

Gael, S., D.L. Grant, and R.J. Richie (1975). Employment test validation for minority and nonminority clerks with worksample criteria. J. Appl. Psychol. 60, 411-419.

Gross, A.L., and Wen-Huey Su (1975). Defining a "fair" or "unbiased" selection model: a question of utilities. J. Appl. Psychol. 60, 345-51.

Irvine, S.H. (1968). Gibt es Kulturunabhängige Tests? - Faktorielle Untersuchungen des Progressiven Matrizentest von Raven in mehreren afrikanischen Kulturen, in: Ingenkamp, K., and Marsolek, T. Möglichkeiten und Grenzen der Testanweisung in der Schule Weinheim: Beltz.

Irvine, S.H. (1969). Figural tests of reasoning in Africa. Int. J. Psychol. 4, 217-228.

Jahoda, G. (1977). In pursuit of the emic-etic distinction: can we ever capture it, in Y.H. Poortinga (ed.) Basic problems in cross-cultural psychology. Amsterdam, Swets & Zeitlinger, 55-63.

Katzell, R.A. and F.J. Dyer (1977).Differential validity revived. J. Appl. Psychol. 62, 137-145.

Kerlinger, F.N. and E.J. Pedhazur (1973). Multiple regression in behavioral research. Holt, Rinehart and Winston Inc.

Kirkpatrick, J.J., R.B. Ewen, R.S. Barret, and R.A. Katzell (1968). Testing and fair employment. New York, New York University Press.

Lord, F.M. (1977). A study of item bias, using item characteristic curve theory, in: Y.H. Poortinga (ed.) Basic problems in cross-cultural psychology. Amsterdam, Swets & Zeitlinger, 19-29.

Mellenbergh, G.J. (1972). An investigation of the applicability of the Rasch model in different cultures, in: Cronbach, L.J. and Drenth, P.J.D., Mental tests and cultural adaptation. Den Haag Mouton.

Mokken, R.J. (1970). A theory and procedure of scale analysis. Den Haag Mouton.

Novick, M.R. and N.S. Petersen (1976). Towards equalizing educational and employment opportunity. J. Educ. Meas. 13, 77-88.

O'Connor, E.J., K.N. Wexley, and R.A. Alexander (1975). Single group validity; Fact or fallacy? J. Appl. Psychol. 60, 352-355.

Petersen, N.S. and M.R. Novick (1976). An evaluation of some models for culture-fair selection. J. Educ. Meas. 13, 3-29.

Plomp, Tj. (1972). Een individueel studiesysteem voor de eerste jaars wiskunde. Technische Hogeschool Twente, Nederland.

Poortinga, Y.H. (1971). Cross-cultural comparison of maximum performance tests: some methodological aspects and some experiments with simple auditority and visual stimuli, in: Psychologica Africana, Monograph, Supplement no. 6.

Poortinga, Y.H. (1975). Limitations on inter cultural comparison of psychological data. Ned. Tijdschrift voor de Psychologie, 30, 23-29.

Poortinga, Y.H. and B.I.M. Foden (1975). A comparative study of curiosity in black and white South African students. Psychologica Africana, Monograph, Supplement no. 8.

Schmidt, F.L., J.G. Berner and J.E. Hunter (1973). Racial differences in validity of employment tests: Reality or illusion? J. Appl. Psychol. 58, 5-9.

Thorndike, R.L. (1971). Concepts of culture-fairness. J. Educ.Meas. 8, 63-70.

Vernon, P.E. (1969). Intelligence and Cultural Environment. London, Methuen.